Kabayaki Basic for Linux

取扱説明書 Version 4.1.0

第 1 版　2014 年 6 月 23 日

はじめに　1

Kabayaki の管理とは　1
　コンテンツとインデックス　1

第 1 章　管理画面の構成　3

管理画面の表示　3

第 2 章　サーバー関連メニュー　5

サーバー環境情報　5
単語一覧　7
　単語の検索・一覧表示　7
単語追加　9
単語の一括登録 / 書き出し　11
単語の修正　13
単語の削除　14
MeCab のユーザー辞書の定義追加　15

第 3 章　インデックス関連メニュー　17

インデックス一覧　17
インデックスの追加　19
インデックスの修正　21
コンテンツ設定　22
フィルタ設定　24
チューニング設定　27
ログ一覧　31
Web スパイダ　33
　基本設定　33
　拡張設定　34

第 4 章　検索および検索結果画面　37

検索画面の表示　37
検索方法　38
　検索式　38

付録 A　文書フィルタと
プロパティ検索詳細　43

文書フィルタ　43
　対応文書　43

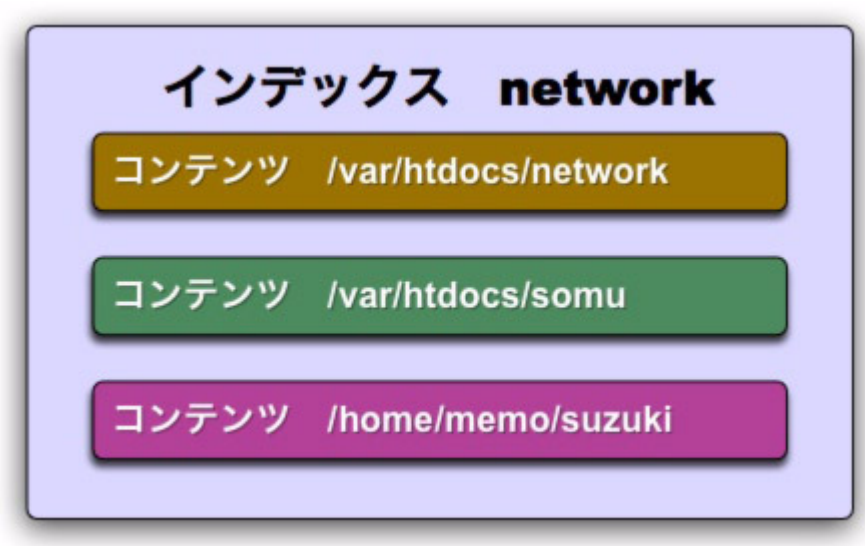
プロパティ検索　45
OLE オブジェクト検索　45

はじめに

Kabayaki の管理とは

コンテンツとインデックス

Kabayaki は、コンテンツとインデックスという枠組みを使用して文書を検索します。コンテンツとインデックスとは次のような関係になります。



上の図では、Kabayaki を運用するシステム上のファイルシステムでは、/var/htdocs/network、/var/htdocs/somu、/home/suzuki/memo という別々の場所で管理している社内ネットワーク関連文書を、network という 1 つのインデックスにまとめて管理しています。

第 1 章　管理画面の構成

管理画面の表示

Kabayaki の管理には、Microsoft Internet Explorer や Firefox などの Web ブラウザを使用します。管理画面を表示させるには、Web ブラウザに次のように URL を入力します。

http:// *ホスト名* /kabayaki/cgi-bin/admin/rc.cgi

「*ホスト名*」の部分には、Kabayaki をインストールしたコンピュータの名前を入力します。

たとえば、インストールしたホストが search.timedia.co.jp ならば、次の URL になります。

http://search.timedia.co.jp/kabayaki/cgi-bin/admin/rc.cgi

Kabayaki が正しくインストールされていると、管理画面が表示されます。

Kabayaki をインストールした状態での画面イメージ :

管理画面左側の一覧をメインメニューと呼びます。メインメニューは、サーバー関連メニューとインデックス関連メニューから構成されます。

サーバー関連メニューでは、Kabayaki をインストールしたホスト全体に関わる情報の設定や表示の種類を選べます。インデックス関連メニューでは、検索に必要なインデックスの情報を設定する機能を選べます。メインメニューは、機能を選んで画面が切り替わっても、常に表示されています。

Kabayaki をインストールして、最初に Kabayaki 管理画面を表示させたときは、サーバー関連メニューの機能とインデックス関連メニューのインデックス一覧しか選ぶことができません。その他の一覧や設定を選ぶには、後述する手順でインデックスを作成する必要があります。

各設定画面に共通で表示されるものには、さらに次のものがあります。

「?」(ヘルプ) ボタン

ページ右上に表示されます。クリックすると、各設定画面のオンラインヘルプが表示されます。

「インデックス選択」

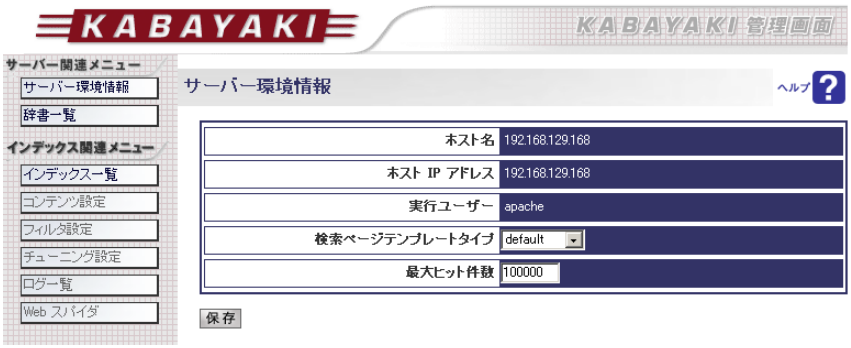
インデックス関連メニューを選択したときのみ表示されます。ページの右上の方に表示されているインデックス名の右の矢印をクリックすると、インデックス名の一覧がメニュー表示されます。メニューからインデックスを選択すると、表示および操作の対象となるインデックスが変更されます。

第 2 章　サーバー関連メニュー

この章では、「サーバー関連メニュー」に分類されている機能について説明します。

サーバー環境情報

Kabayaki をインストールしたホストに関する情報を表示します。また、検索および検索結果の画面のデザイン、最大ヒット件数を変更できます。管理画面の左側に表示されるメニューの「サーバー環境情報」ボタンをクリックすると、この「サーバー環境情報」画面が表示されます。



画面に表示されている情報は以下の通りです。

ホスト名

Kabayaki が動作しているホストの名前が表示されます。環境変数 SERVER_NAME を参照しています。SERVER_NAME が設定されていないと、「不明」と表示されます。

ホスト IP アドレス

Kabayaki が動作しているホストの IP アドレスが表示されます。これは、環境変数 SERVER_ADDR を参照しています。SERVER_ADDR が設定されていない場合は、「不明」と表示されます。

実行ユーザー

Kabayaki 管理画面を実行しているプロセスの実行ユーザー名が表示されます。

検索ページテンプレートタイプ

検索と検索結果の画面の外観を選択できます。default はインストールした直後と同じ画面になります。

最大ヒット件数

検索実行時の最大ヒット件数を指定できます。既定値は 10 万件です。

あるキーワードの検索結果が、この設定値を超えた件数になると、その検索キーワードは無視されます。たとえば、「。」と「検索」で AND 検索を実行し、「。」を含む文書の数がここで指定した「最大ヒット件数」よりも多かった場合には、「ヒット数が多すぎるので無視した検索キーワードが存在します」というエラーメッセージが表示され、「検索」というキーワード単独で検索を実行したのと同じ結果が一覧表示されます。

「保存」ボタン

「保存」ボタンをクリックすると、検索ページテンプレートタイプの設定が保存されます。

検索テンプレートが使用された検索結果画面

default :

単語一覧

管理画面の左側に表示されるメニューの「単語一覧」ボタンをクリックすると、「単語一覧」画面が表示されます。

単語の検索・一覧表示

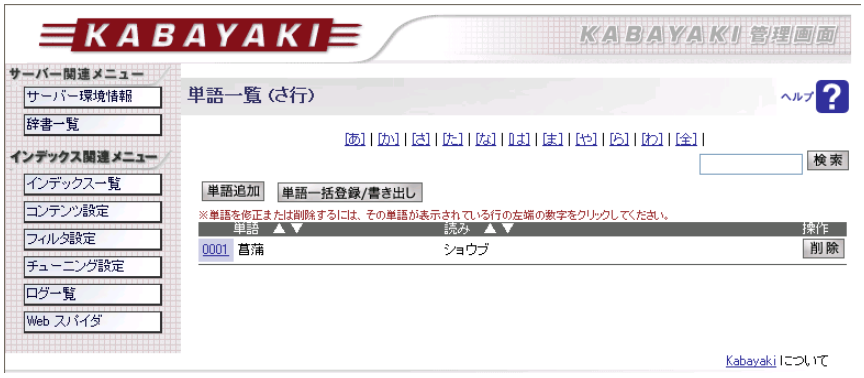
単語の登録、修正、削除の際に登録済みの単語を確認するには、辞書を検索する機能を利用すると便利です。

登録されている全ての単語を一覧表示

「単語一覧」の画面の「[全]」と書いてあるリンクをクリックします。

登録した単語の頭文字で検索

「単語一覧」の画面から、単語の読みの頭文字のリンクをクリックします。たとえば、「菖蒲」を探しているときは「さ」を選択します。



単語で検索
右上段にある検索ボックスに、単語や単語の読みの一部または全部を入力して検索します。

該当する単語があれば、一覧に表示されます。

単語追加

ここでは、「災害復旧ソフトウェア」という単語を登録する場合を例にとって説明します。この「災害復旧ソフトウェア」という語を検索画面のキーワードとして指定すると、通常の状態では「災害」「復旧」「ソフトウェア」という 3 つの単語が別々に並んでいるものと扱われるため、「二次災害」「復旧への道」「ソフトウェア購入」などが含まれた文書があると、それも検索結果に含まれてしまいます。検索精度の向上のために「災害復旧ソフトウェア」をひとつの単語として認識させたいならば、この語を新しく辞書に追加する必要があります。

単語を新規登録するには、単語一覧の左上の方にある「単語追加」ボタンをクリックします。

この画面では、以下の入力項目が存在します。

- 単語
- 読み

「単語」には、Kabayaki のシステムに 1 つの単語として認識させたい語を入力します。今回は「災害復旧ソフトウェア」という単語を登録するので、「災害復旧ソフトウェア」と入力します。

「読み」には、単語の読みを片仮名で入力します。今回は「災害復旧ソフトウェア」という単語を登録するので、読みのフィールドには「サイガイフッキュウソフトウェア」と入力します。

ここで入力できるのはカタカナのみです。カタカナ以外が含まれているとエラー画面が表示されますので、画面の指示に従って再度入力してください。

全ての項目を入力したら、「登録」のボタンをクリックします。

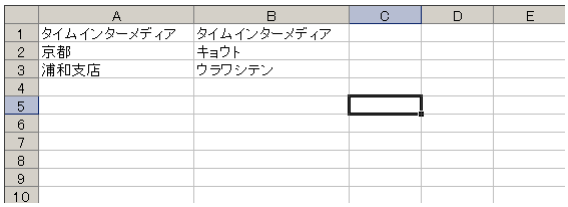
登録された単語は、次回の再構築後に有効になります。

単語の一括登録 / 書き出し

一度にたくさんの単語を登録するときに、「単語追加登録」画面から 1 つずつ登録すると時間がかかります。Kabayaki では、Microsoft Excel で作成した単語の一覧を、「単語一括登録」画面からアップロードすることにより、複数の単語を一度に登録することが可能です。

まず、Microsoft Excel を使用して、一括で登録する単語の一覧を以下の書式で作成します。

単語 , 読み



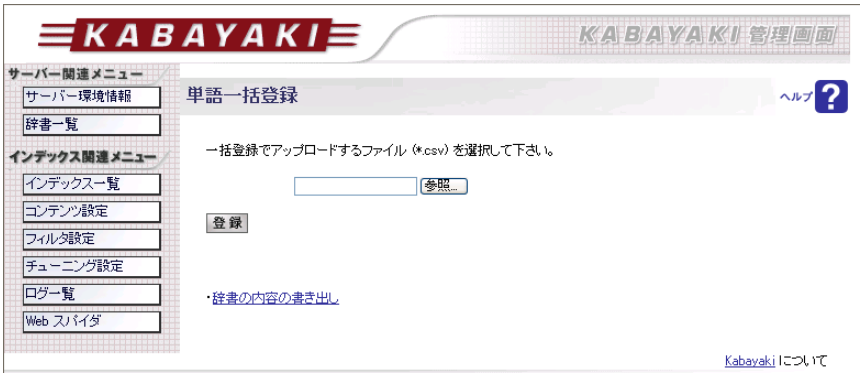
	A	B	C	D	E
1	タイムインターメディア	タイムインターメディア			
2	京都	キョウト			
3	浦和支店	ウラワシテン			
4					
5					
6					
7					
8					
9					
10					

Microsoft Excel の操作に関する詳細は、製品付属の取扱説明書を参照してください。

読みに使用できるのは、全角のカタカナのみです。

注) Microsoft Excel でファイルを保存する際には、『CSV 形式』で保存してください。CSV 以外の形式で保存したファイルを、Kabayaki 辞書管理ツールの一括登録に使用することはできません。

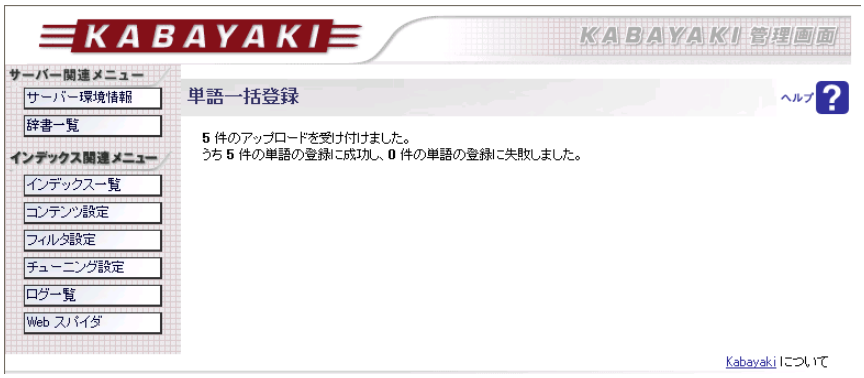
次に、Kabayaki 管理画面の「単語一覧」の左上の方 (「単語追加」ボタンの隣) にある「単語一括登録 / 書き出し」ボタンをクリックします。



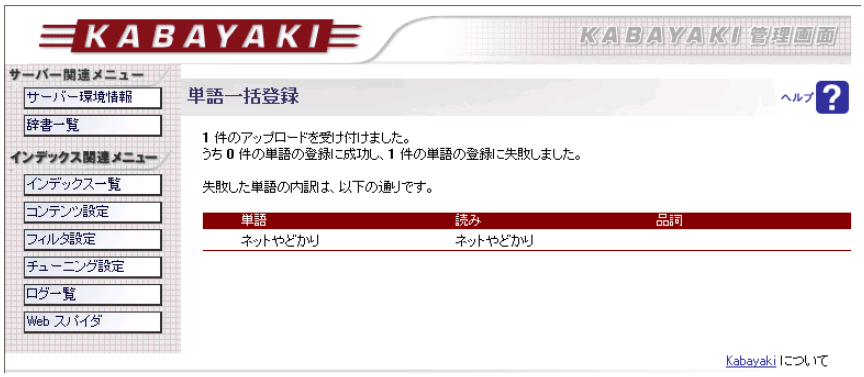
表示された「単語一括登録」画面から、作成済みの CSV ファイルをアップロードします。

これで単語の一括登録の処理が完了し、実行結果が表示されます。画面には、単語の読み取り件数、登録の成功件数と失敗件数が表示されます。登録を失敗した単語があれば、登録を失敗した単語が一覧表示されます。登録に失敗した単語に関しては、CSV ファイルを修正して再度登録してください。

全ての単語が登録できた例：



登録がエラーになった例：



辞書の書き出し

このリンクをクリックすると、登録されている単語の一覧を、CSV 形式のファイルで取得することができます。

この機能を使用して書き出されたファイルを Microsoft Excel で読み込んで、編集することもできます。

単語の修正

「単語の一覧」画面から、登録内容の更新ができます。

ここでは、「ショウブ」という読みで登録していた「菖蒲」という単語の読みを、「アヤメ」に修正する場合を例にとって説明します。

まず、単語の一覧表示で「菖蒲」を表示させます。表示させる方法は、前述の「単語の検索・一覧表示」を参照してください。

「菖蒲」と表示されている行の、左端の数字のリンクをクリックします。

次に表示される画面で修正したい項目を変更できます。この例では、「読み」の「ショウブ」を「アヤメ」に変更してから、「登録」ボタンをクリックします。

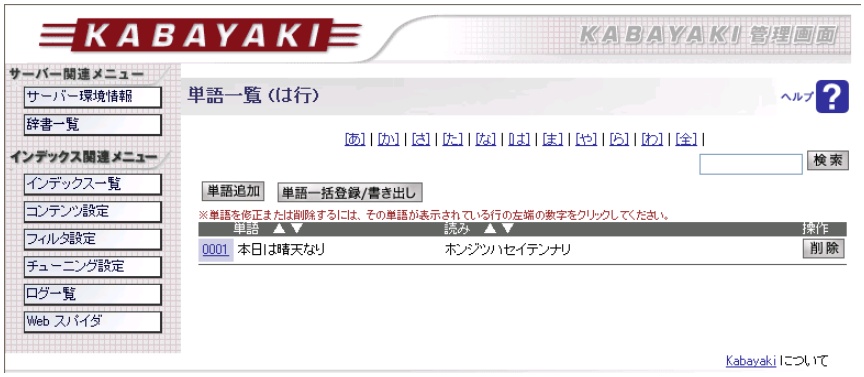
単語の登録に失敗した場合は、画面に表示されるエラーメッセージを参照して、操作をやり直してください。

単語の削除

「単語一覧」画面から、登録単語を削除できます。

たとえば、誤って追加してしまった「本日は晴天なり」という単語を削除したい場合は、以下のようにします。

まず、単語の一覧表示で「本日は晴天なり」を表示させます。表示させる方法は、前述の「単語の検索・一覧表示」を参照してください。



「本日は晴天なり」と表示されている行の、右端の「削除」ボタンをクリックします。

削除の確認を求める画面が表示されます。削除してよければ「削除」ボタンをクリックします。

単語の削除に失敗した場合は、画面に表示されるエラーメッセージを参照して、操作をやり直してください。

MeCab のユーザー辞書の定義追加

Kabayaki 管理画面の「辞書一覧」を使用して単語が登録されたユーザー辞書を、検索用インデックス生成の際に正しく参照させるためには、C:¥Program Files¥mecab¥etc¥mecabrc に次の行を指定する必要があります。

```
userdic = /var/kabayaki/dic/timecab.dic
```

注意：ユーザー登録された単語がない状態で、この行が定義されていると、インデクシングが異常終了します。登録した単語を全て削除した場合や、初期状態の単語が登録されていない状態では、この行を削除またはコメントアウトしておく必要があります。

第３章　インデックス関連メニュー

インデックス一覧

管理画面の左側に表示されるメニューの「インデックス一覧」ボタンをクリックすると、「インデックス一覧」画面が表示され、登録されているインデックスが一覧表示されます。

インストール直後等の、登録されているインデックスが１つも存在しないときは、

新しくインデックスを作成してください。

と表示されます。なお、インデックスについては、『はじめに』の「Kabayaki の管理とは コンテンツとインデックス」を参照してください。

「インデックス追加」ボタン

入力されている内容で、インデックスを新規に登録します。

インデックス一覧の表示は、「内部名」のアルファベット順です。
一覧で表示される項目は次の通りです。

表示名

表示名は、他の Kabayaki 管理画面や検索結果画面でも表示される名前です。クリックすると、そのインデックスの「コンテンツ設定」画面が表示されます。

内部名

内部名は、Kabayaki が内部的に使用する名前です。検索結果画面の右側に表示されるインデックス一覧は、この内部名の順に表示されます。表示されている内部名をクリックすると、そのインデックスの「コンテンツ設定」画面が表示されます。

コンテンツ種別

インデックスのコンテンツの種類が表示されます。ここに表示される内容は、インデックス作成時の指定および「コンテンツ設定」画面での設定に応じて決まります。

コンテンツが未設定のインデックスでは「-」が表示され、コンテンツにローカルパス上のファイルが指定されていると「ファイル」、Web コンテンツに http:// で始まるリモートパスが指定されていると「Web」と表示されます。両方が指定されているときは「ファイル /Web」という表示になります。

状態

インデクシング中には、ここに「処理中」と表示されます。それ以外のときは「-」が表示されます。

操作

修正ボタンを押すと、そのインデックスの表示名を変更するための画面に移動します。

削除ボタンを押すと、そのインデックスを削除するための画面に移動します。

インデックスの追加

「インデックス一覧」画面の「インデックス追加」ボタンをクリックすると、「新規インデックス追加フォーム」が表示されます。

項目は次の通りです。

内部名

インデックス識別のために Kabayaki が内部的に使用する名前を入力します。検索結果画面の右側に表示されるインデックス一覧は、この内部名の順に表示されます。半角小文字の英数字とアンダースコア (_) のみが入力できます。

0123456789 abcdefghijklmnopqrstuvwxyz _

表示名

他の Kabayaki 管理画面や検索結果画面で表示されるインデックス名を入力します。機種依存文字や登録外字、半角カタカナは使用できません。また、半角の # ! & < > % ' ″ | ( ) ¥ や空白文字も指定できません。

内部名、表示名ともに、入力できる文字数の制限を超えて入力することはできません。

インデックスの数は 64 個まで作成・検索可能です。なお、インデックスの内部名の長さや、Web サーバー、Web ブラウザによって、検索可能なインデックスの数は 64 個よりも少なくなることがあります。インデックスの内部名は検索時の GET パラメータとして利用されるため、作成するインデックスの数が多くなる場合は、なるべく短い名前にすることをお勧めします。

また、含まれる文書数については特に数値的な上限は設定されていませんが、文書の総数は 10 万文書程度までを目安にして、インデックス分割により文書を分散させることをお勧めします。

「登録」ボタン

入力されている内容で、インデックスを新規に登録します。

注意

インデクシングが行なわれている最中に、インデックス追加や削除を実行しないでください。このような操作をしますと、以後、インデクシングや検索が正しく動作しなくなることがあります。

インデックスの修正

「インデックス一覧」画面の、各インデックスの列の右にある「修正」ボタンををクリックすると、「インデックス表示名修正フォーム」が表示されます。

項目は次の通りです。

内部名

インデックス識別のために Kabayaki が利用する名前が表示されます。表示のみで変更はできません。。

表示名

他の Kabayaki 管理画面や検索結果画面で表示されるインデックス名を入力します。機種依存文字や登録外字、半角カタカナは使用できません。また、半角の # ! & < > % ' ″ | ( ) ¥ や空白文字も指定できません。

内部名、表示名ともに、入力できる文字数の制限を超えて入力することはできません。

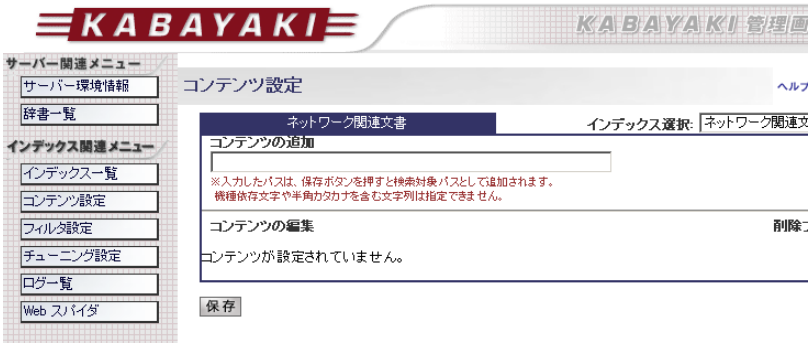
「登録」ボタン

入力されている内容で、Kabayaki のインデックスの表示名を変更します。これにより、Kabayaki 管理画面に表示されるインデックスの名前と、検索結果画面に一覧表示されるインデックスの名前が変更されます。

コンテンツ設定

インデックスに設定されるコンテンツ (検索対象の文書がある場所) を、追加・設定・削除します。コンテンツについては、『はじめに』の「Kabayaki の管理とは コンテンツとインデックス」を参照してください。

管理画面の左側に表示されるメニューの「コンテンツ設定」ボタンをクリックするか、「インデックス一覧」画面で一覧表示されている表示名または内部名のリンクをクリックすると、この「コンテンツ設定」画面が表示されます。



画面に表示されている情報は以下の通りです。

コンテンツの追加

入力フィールドに、インデクシングを実行したい (検索対象にしたい) 文書が含まれるディレクトリのパスを、絶対パスで入力します。検索の対象となるファイルは、指定されたディレクトリ以下の全てのファイルとなります。

このフィールドに機種依存文字や半角カタカナを含む文字列を指定することはできませんので、注意してください。

sansyo パッケージが導入されていて、Web スパイダが動作しているシステムでは、ここに http:// で始まるリモートパスを指定することも可能です。

コンテンツの編集

コンテンツの一覧が表示されます。インデックスが作成された直後の時点では、設定されているコンテンツは存在しません。

コンテンツの追加フィールドには、検索の結果に表示される文書の場所を指定します。コンテンツ (検索対象) としてローカルパスが指定された直後は、「閲覧時の URL」欄には「http:// ローカルパス /」のようにローカルパスがそのまま入るため、検索を実行したユーザーの Web ブラウザから文書を参照可能な URL にするための変更が必要になることがあります。たとえば、コンテンツが存在するディレクトリが

```
/var/www/html
```

に設定されている Web サーバー search.timedia.co.jp の場合は、次のように「閲覧時の URL」を編集します。

http://var/www/html/yamada → http://search.timedia.co.jp/yamada

http://var/www/html/yamada_syanaihi → http://search.timedia.co.jp/yamada_syanaihi

http://usr/local/apache/htdocs/yamada → http://search.timedia.co.jp/yamada

「削除フラグ」チェックボックス

コンテンツを削除するには、「コンテンツの編集」の一覧の、削除したいコンテンツの右側に表示されている「削除フラグ」チェックボックスをチェックしてから、「保存」ボタンをクリックします。

文書の格納されている場所によっては、Web ブラウザで検索結果を正しく表示させるために、Web サーバーの設定ファイルの編集が必要になることがあります。

Apache HTTP Server では、標準の設定ファイルである httpd.conf を編集します。たとえば、以下のような行を httpd.conf に追加します。

```
Alias    /doc/network     "/usr/local/doc/network"
```

コンテンツの追加で、異なるディレクトリパスを設定するたびに、上記のような Alias ディレクティブを追加していきます。

Alias を追記したら、Apache HTTP Server を再起動して、httpd.conf の変更を反映させます。

```
# apachectl graceful
```

または

```
# httpd restart
```

フィルタ設定

ファイルの拡張子で示されるファイル形式やサブディレクトリを指定して、検索の対象にするコンテンツを選別 (フィルタリング) することができます。フィルタ設定を上手に利用することによって、無駄なファイルのインデクシングを回避し、インデクシング時間やホストの資源を節約することができます。設定項目の優先順位が低い順に並べると、次の通りになります。

優先順位　低 ⟷ 高　優先順位

検索対象ファイル ＜ 検索対象外ファイル拡張子 ＜ 検索対象外パス

管理画面の左側に表示されるメニューの「フィルタ設定」ボタンをクリックすると、この「フィルタ設定」画面が表示されます。

検索対象ファイル

検索の対象にしたいファイル名の拡張子を選び、チェックボックスをチェックします。

- HTML ファイルは拡張子が以下のものを対象とします。
  html、htm、phtml、shtml、html. *英数2 文字*
- Mail/News、man 形式は以下のものを対象とします。
  数字のみ、または、文字 + 数字
- 一太郎は拡張子が以下のものを対象とします。
  jaw、jtd

「その他のファイル」をチェックすると、以下のファイルを除く全てのファイルを検索対象とします。

- アーカイブファイル (*.tar、*.tgz、*.lzh、*.zip)
- Windows システムファイル (*.exe、*.dll)
- Microsoft Visio ファイル (*.vsd)
- Microsoft Project ファイル (*.mpp)
- Microsoft Access ファイル (*.mdb)
- メディアファイル (*.wav、*.wmv、*.wmz、*.swf)
- 画像ファイル (*.psd、*.ai、*.gif、*.png、*.jpg、*.jpeg、*.dib、*.bmp、*.tif、*.tiff)
- # で始まるファイル

初期設定では「その他のファイル」がチェックされているため、ここに挙げられていない .php や .cgi といった拡張子を持つファイルは検索対象となります。「その他のファイル」のチェックをはずすと、拡張子なしのファイルや動的に生成される Web ページの多くが検索対象外となる可能性があるため、注意が必要です。

検索対象外のファイル名

検索の対象にしたくないファイル名の一部または全体をテキストエリア内に、1 行にひとつずつ記述します。*.sit、*.c のようにワイルドカードとしての * を含んだ形式で指定することにより、検索対象外の拡張子を指定することもできます。

例 ）

```
htaccess
*.sit
*.c
```

検索対象外パス

検索の対象にしないディレクトリのパスを指定します。「コンテンツ設定」で指定されているパスの下に存在するが、検索の対象には含めたくないサブディレクトリを指定します。

このフィールドに機種依存文字や半角カタカナを含む文字列を指定することはできませんので、ご注意ください。また、ここで指定できるのはローカルパスのみです。http:// で始まるリモートパスは、「Web スパイダ」の「巡回除外パス」で設定してください。

「保存」ボタン

「保存」ボタンをクリックすると、入力されているフィルタ設定を保存します。

注意

Kabayaki Basic 2.1.0 以前のバージョンからアップグレードした場合には、アップグレード直後に各インデックスの「フィルタ設定」画面の設定を確認し、「保存」ボタンを押しておくことをお勧めします。

Kabayaki Basic 2.1.0 では、Office 2007 文書の新規サポートおよびバグ

フィックスのために、設定ファイルの内部形式に変更を加えているためです。

注意 2

テキストおよび HTML、Microsoft Office 文書、PDF、RTF、一太郎、オアシスについては、Kabayaki サーバーに別途ソフトウェアをインストールする必要がありません。それ以外の形式のファイル、bzip2 圧縮ファイル、Pack 圧縮ファイル、Postscript、man、TeX ファイルなどについては、処理するためのソフトウェアがインストール済みの状態でない場合には、追加インストールが必要となることがあります。

チューニング設定

「チューニング設定」画面では、日々変化するコンテンツのインデクシングを効率よく管理するための、インデクシング実行の間隔や時刻の設定、メモリ設定等の設定変更機能を提供しています。

インデクシング間隔

インデクシングを実行する間隔、開始時刻、曜日を設定します。インデックス毎に異なった時刻を設定できます。インデクシングの処理対象となるのは、新規追加または更新されたファイルです。

実行間隔を、「設定しない」、「*n* 日毎」、「毎週、指定された曜日に実行」から選択します。

- 「*n* 日毎」、「毎週、指定された曜日に実行」のどちらかを選択すると、「開始時刻 :」で選択した時刻に処理が開始されます。
- 「*n* 日毎」のラジオボタンを選択したときは、何日毎に実行するかをプルダウンメニューで選択できます。既定値は「1 日毎」で、「1 日毎」から「30 日毎」までが選択可能です。
- 「毎週、指定された曜日に実行」を選択すると、「日」から「土」までの曜日選択が有効になります。曜日は複数指定できます。指定省略時は「日」曜日のみが選択されます。

再構築間隔

インデックスの再構築を実行する間隔、開始時刻、曜日を設定します。インデックス毎に異なった時刻を設定できます。処理対象となるのは、全ファイルです。

- 実行間隔を、「設定しない」、「*n* 日毎」、「毎週、指定された曜日に実行」、「*n* 月毎、*n* 日に実行」から選択します。
- 「設定しない」以外を選択すると、「開始時刻 :」で選択した時刻に処理が開始されます。
- 「*n* 日毎」のラジオボタンを選択したときは、何日毎に実行するかをプルダウンメニューで選択できます。既定値は「1 日毎」で、「1 日毎」から「30 日毎」までが選択可能です。
- 「毎週、指定された曜日に実行」を選択すると、「日」から「土」までの曜日選択が有効になります。曜日は複数指定できます。指定省略時は「日」曜日のみが選択されます。
- 「*n* 月毎、*n* 日に実行」のラジオボタンを選択したときは、何か月毎に実行するかと、何日に実行するかを、プルダウンメニューで選択できます。既定値は「1 月毎、1 日に実行」で、月の間隔は 1、2、3、4、6、12 から、日付は 1 日から 31 日までが選択可能です。実行される月は、1 月から数えて *n* 月毎です。(例 :「4 月毎、15 日に実行」を指定すると、1 月 15 日、5 月 15 日、9 月 15 日の年 3 回の実行)
  ※ある月に存在しない日付が指定されていると、その月には再構築が実行されません。31 日を指定する場合などはご注意ください。29 日 (うるう年以外) や 30 日の指定だと 2 月には処理が実行されません。

インデクシング、再構築「手動実行」ボタン

左側の欄に存在する「手動実行」ボタンを押すと、すぐにインデクシングまたは再構築が開始されます。

オプションパックの導入や辞書の追加や変更をしたときなど、インデックスを作り直す必要のあるときは、「再構築間隔」の欄の方にある「手動実行」ボタンをクリックして、インデックスを再構築します。

注意 1

インデクシングの処理実行中に、インデックス追加や削除の操作はしないでください。このような操作をすると、以後、インデクシングや検索が正しく動作しなくなることがあります。

注意 2

Red Hat Enterprise 4.0 を使用している場合は、最初のインデックスを新規追加した直後に crond を再起動しないと、間隔の設定が有効になりませんので注意してください。

検索結果設定

製品版にのみ表示される項目で、検索結果画面に表示される、検索で見つかったファイルの一部の表示形態を選択します。

「ファイルの先頭を表示」では常にファイルの先頭を表示します。「検索語の周辺文章を表示」では、検索語が検索画面に現れるようにするために、ファイルの文章中から検索語の周辺を切り出して表示します。

検索結果設定の変更は、再構築実行後に有効になります。

使用メモリーサイズの上限値

使用メモリーサイズを 1000 バイト単位で指定します。指定可能な最大値は 10000000 (10G) です。

この値を大きくすると、インデックス書き出しの頻度が下がるため、インデクシングにかかる時間が短縮されることがあります。ただし、ホストの物理メモリーが少ない場合などは、かえって遅くなることもあるため注意が必要です。同時にインデクシングを実行するインデックスの使用メモリーサイズの合計が実メモリー以上の値になるような設定は避けてください。

ファイルサイズの上限値

インデクシングの対象となるファイルの大きさを 1000 バイト単位で指定します。指定可能な最大値は 100000 (100M) です。

この値よりも大きなファイルはインデクシングの対象にはなりません。意図しない検索漏れが発生しないように注意して設定する必要がありますが、マシンの処理能力を超えた設定はインデクシングの異常終了の原因となることがあります。後述の「設定に際しての注意」を参考に、ハードウェア仕様および他に実行する処理の負荷を考慮して設定してください。

テキストの上限値

インデクシングの対象となるファイルのテキスト部分の大きさを 1000 バイト単位で指定します。指定可能な最大値は 50000 (50M) です。

この値よりも大きなテキストのサイズのファイルはインデクシングされません。マシンの処理能力を超えた設定はインデクシングの異常終了の原因となることがあります。後述の「設定に際しての注意」を参考に、ハードウェア仕様および他に実行する処理の負荷を考慮して設定してください。

設定に際しての注意

「使用メモリーサイズの上限値」、「ファイルサイズの上限値」、「テキストの上限値」、「単語長の上限値」といった上限値を増やすにあたっては、それに応じたハードウェア性能が、正常動作のためには必要です。

安定稼動のためには、メモリーが 2G ならば、「ファイルサイズの上限値」は 25000 程度、「テキストの上限値」は 3000 程度を目安に、ログを監視しながら段階的に増やしていく運用をお勧めします。

インデクシングに際しての注意

Kabayaki は、1 つのインデックスに対して、同時に複数のインデクシングプロセスを実行することを禁止しています。そのため、インデクシング実行中に同じインデックスに対してインデクシングを実行しようとすると、以下のようなメッセージがログに出力されることがあります。

> "... not executed, because */vardir*/kabayaki/targets/*idxname*/lock exists."

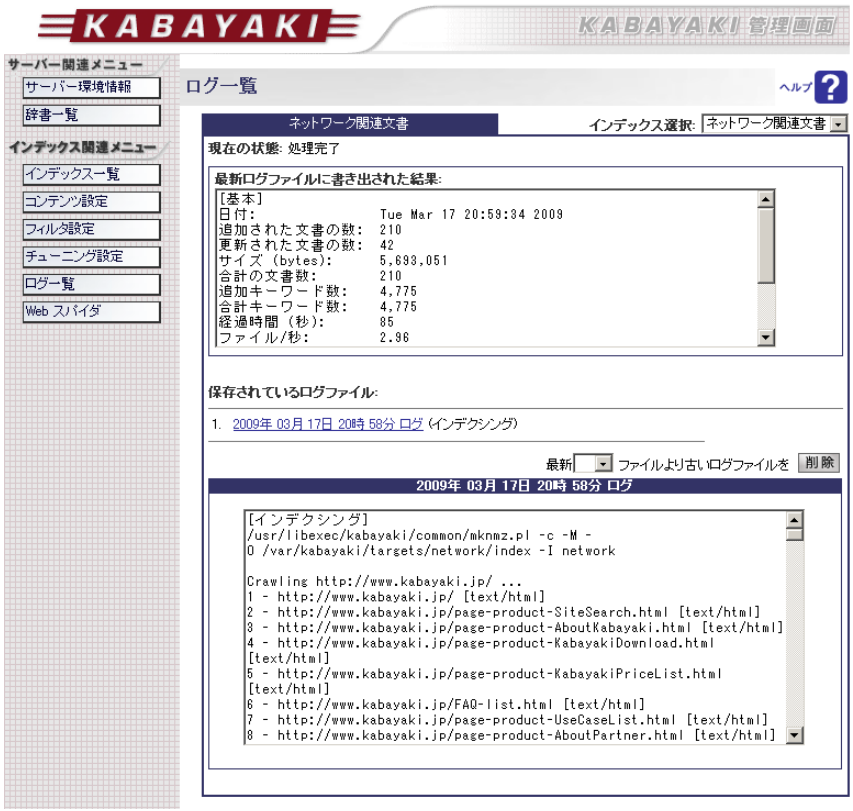
vardir には、Red Hat なら /var が入ります。*idxname* に入るのはインデックスの内部名です。

外部的または内部的な要因によってKabayakiのプロセスが意図しない異常終了を起こしたときも、該当するインデックスのログ一覧には上記のようなメッセージが表示されます。このような場合は、ロックファイルがファイルシステムに残ったままになっていることにより、それ以後のインデクシングができなくなることがあります。この状態から再びインデクシング実行を可能にするためには、以下の手順で作業してください。

1. 管理画面の左側に表示されるメニューの「ログ一覧」ボタンをクリックして、ログ一覧を表示させます。
   (ログについては、次の節の「ログ一覧」を参照してください)。
2. 「ログ一覧」で上記のメッセージを確認して、ロックファイル /*vardir*/kabayaki/targets/*idxname*/lock を手動で削除します。
3. /*vardir*/kabayaki/index/*idxname*/NMZ.lock2 ファイルが存在していたならば、そちらも削除します。
4. 「チューニングの画面」で、インデクシングまたは再構築の手動実行ボタンを押して、処理が正常に開始されるかどうかを確認します。

ログ一覧

インデクシングに関するログを表示します。ホストのシステム管理者は、ログを定期的に参照することによって、処理状況を確認できます。管理画面の左側に表示されるメニューの「ログ一覧」ボタンをクリックすると、次のような画面が表示されます。



現在の状態 :

インデクシングの処理中だと、

- 「インデクシング処理中」

と表示されます。処理が完了していると

- 「処理完了」

と表示されます。「処理完了」のときはその下に表示される「最新ログファイルに書き出された結果 :」で、処理結果を参照できます。

保存されているログファイル :

インデクシングの実行結果のログファイルが、日付の新しいものから一覧表示されます。日付の部分をクリックすると、画面下部のテキスト領域に詳細なログが表示されます。

詳細表示される項目には次の情報があります。

インデクシング

- 日付
- 追加された文書数
- 削除された文書数
- サイズ
- 更新された文書数
- 合計の文書数
- 追加キーワード数
- 合計キーワード数
- わかち書き
- 経過時間
- ファイル / 秒
- 秒 / ファイル
- システム
- Namazu
- エラー、警告、詳細

最新 *n* ファイルより古いログファイルを削除

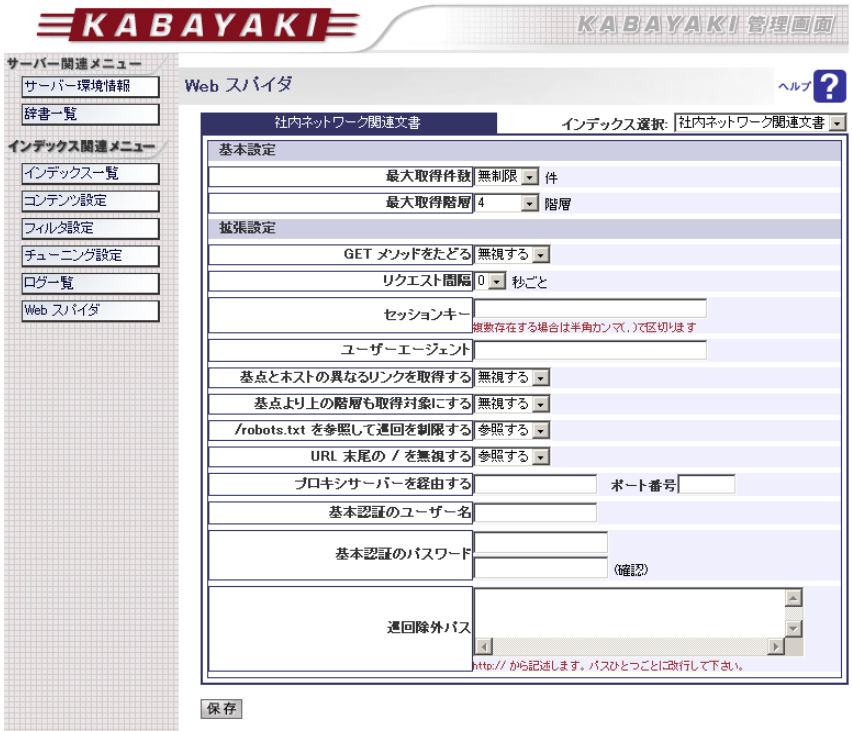
古いログファイルを自動的に削除することが可能です。初期設定ではログは自動的に削除されません。

プルダウンメニューから残したいログの数を選択して「削除」ボタンを押すと、その時点で指定された数のログファイルだけ残し、古いログファイルが削除されます。また、それ以降のインデクシングの度に古いファイルから順に削除され、常に指定された数のログファイルだけが残るようになります。

数値指定なし (空白) の場合はログの個数を制限しないため、自動削除は実行されません。

Web スパイダ

Web スパイダは、Web サイトを巡回しコンテンツを収集する機能で、製品版 Kabayaki でのみ提供されます (GPL 版の Kabayaki では、この機能は提供されません)。Web スパイダでは、様々な Web サイトの仕組みに対応し、取り込んだコンテンツを検索できるようにするための、きめ細かな設定が可能になっています。



基本設定

最大取得件数

収集するコンテンツの数を制限します。初期設定は「無制限」で、収集するコンテンツの数は制限されていません。

最大取得階層

探索するハイパーリンクの階層の数を制限します。コンテンツ設定で入力した URL を基点とし、そこからハイパーリンクをたどる回数を指定します。初期設定は「4」です。URL に含まれる / の数とは関係ありません。

注意

「最大取得件数」と「最大取得階層」の両方を、「無制限」に設定することは避けてください。Kabayaki サーバーのみでなく、Web サーバー等に過大な負荷をかける原因となる可能性があります。

拡張設定

GET メソッドをたどる
URL に表れる ? より右側の & で区切られた = をはさんだキーと値の組み合わせ (クエリー) を、URL の一部とみなすかどうかを設定します。初期値は「無視する」ですので、クエリーを URL の一部とはみなさずに探索します。

リクエスト間隔
Web スパイダからサーバーへコンテンツの取り出しを要求する間隔を秒数で指定します。

セッションキー
セッションキーを取り除いたものを URL とするかどうかを設定します。初期値は空白です。セッションキーにあたる文字列を入力すると、それを取り除いたものを URL と見なして探索します。

ユーザーエージェント
Web サーバーへ送信するユーザーエージェント情報を文字列で設定します。初期値は空白です。

基点とホストの異なるリンクを取得する
異なる Web サーバーのコンテンツも探索するかどうかを設定します。初期設定は「無視する」で、基点となる URL から Web サーバーを越えた探索をしません。

基点より上の階層も取得対象にする
コンテンツ設定で入力した URL を遡って探索するかどうかを設定します。初期設定は「無視する」で、ハイパーリンクや CGI で URL が遡ったコンテンツは探索しません。

なお、特定の URL が基点より上と判断されるかどうかは、基点の URL の指定方法に依存します。たとえば http://www.example.com/file.html は、コンテンツ設定画面で http://www.example.com/dir/ を基点に指定した場合は、http://www.example.com/file.html は基点より上と判断され、http://www.example.com/dir と指定した場合は基点と同じ階層と判断されます。

/robots.txt を参照して巡回を制限する
robots.txt の内容に従って巡回を制限するかどうかを設定します。初期設定は「参照する」で、robots.txt の内容を遵守して探索します。

URL 末尾の / を無視する
一部のサイトではコンテンツの URL 末尾に「/」(スラッシュ) がついていることがあります。これを取り除いてインデクシングの対象とします。

プロキシサーバーを経由する
プロキシ (HTTP PROXY) を経由しないと Web コンテンツを取得できないネットワーク環境にいる場合、経由させるプロキシサーバーのホスト名と使用するポート番号を指定します。

基本認証のユーザー名および基本認証のパスワード

基本認証 (BASIC 認証) のあるコンテンツを閲覧する時に必要なユーザー名とパスワードを設定します。初期値は空白で、基本認証の必要なコンテンツは探索しません。

巡回除外パス

Web コンテンツを取得しなくてもよい URL を指定します。除外したい URL が複数あるときは、1 つの URL につき 1 行ずつ入力します。

第 4 章　検索および検索結果画面

検索画面の表示

インデックス作成後に実際に検索するには、Microsoft Internet Explorer や Firefox などの Web ブラウザに以下のような URL を入力して、検索画面を表示させます。

http:// *ホスト名* /kabayaki/

「*ホスト名*」の部分には、Kabayaki をインストールしたコンピュータの名前を入力します。

たとえば、インストールしたホストが search.timedia.co.jp ならば、次の URL になります。

http://search.timedia.co.jp/kabayaki/

Web ブラウザには、次のような検索画面が表示されます。

検索画面 :

「表示件数」には検索結果画面に一度に表示する件数、「ソート」には検索結果画面上での並べ替え順を指定します。

検索文字列を指定して「検索」ボタンを押すと、検索結果画面が表示されます。

検索結果画面 :

検索方法

検索式

単一単語検索

調べたい単語を 1 つ指定するだけのもっとも基本的な検索手法です。
例 :

namazu

AND 検索

ある単語とある単語の両方を含む文書を検索します。検索結果を絞り込むのに有効です。3 つ以上の単語を指定することも可能です。単語と単語の間に and または & を挿みます。
例 :

Linux and Netscape

and または & は省略できます。単語を空白で区切って羅列するとそれらの語すべてを含む文書を AND 検索します。

OR 検索

ある単語とある単語のどちらかを含む文書を検索します。3 つ以上の単語を指定することも可能です。単語と単語の間に or または | を挿みます。
例 :

Linux or FreeBSD

NOT 検索

ある単語を含み、ある単語を含まない文書を検索します。3 つ以上の単語を指定することも可能です。単語と単語の間に not または ! を挿みます。
例 :

Linux not UNIX

グループ化

AND 検索、OR 検索、NOT 検索を括弧でグループ化できます。括弧の両隣には空白を入れる必要があります。
例 :

(Linux or FreeBSD) and Netscape not Windows

部分一致検索

部分一致検索には前方一致、中間一致、後方一致の 3 種類があります。

- 前方一致検索
  inter* (inter から始まる単語を含む文書を検索)
- 中間一致検索
  text (text を内包する単語を含む文書を検索)
- 後方一致検索
  *net (net で終わる単語を含む文書を検索)

フィールド指定の検索

Subject:、From:、Message-Id: といったフィールドを指定して検索する手法です。特に Mail/News のファイルを扱う際に効果を発揮します。

例 :

- +subject:Linux
  (Subject: に Linux が含まれる文書)
- +subject:"GNU Emacs"
  (Subject: に GNU Emacs が含まれる文書)
- +from:foo@bar.jp
  (From: に foo@bar.jp が含まれる文書)
- +message-id:<199801240555.OAA18737@foo.bar.jp>
  (Message-Id を指定)

特記事項

- いずれの検索方法でもアルファベットの大文字・小文字の区別 はしません。
- 日本語の複合語は形態素単位に分割し、それらをフレイズ検索します。分割は不適切に行なわれることがあります。
- JIS X 0208 (いわゆる全角文字) の英数字と記号の一部 (ASCII と重複しているもの) は ASCII (いわゆる半角文字) として処理されます。

- JIS X 0201 (いわゆる半角カタカナ) は、全角カタカナとして処理されます。全角カタカナでの検索によって、元の文書中に含まれた半角カタカナの単語を検索可能です。
- 記号を含む語の検索ができます。例 : TCP/IP。ただし、記号の処理は完全ではないので TCP and IP のように分割して AND 検索をかけた方が取りこぼしがありません (その代わり余計なファイルまでヒットしてしまう可能性があります)。
- 中間一致・後方一致、正規表現、フィールド指定の検索には少し時間がかかります。
- and、or、not を単語として検索したいときはそれぞれ、"..." と 2 重引用符で、あるいは {...} と中括弧で囲みます。

ソート

標準の Kabayaki パッケージでは、検索結果を、スコア、日付、題名、著者、サイズ、URI の順で並べ替えることができます。既定値はスコア順でのソートです。

スコア

- 基本的には、指定された検索文字列の登場回数が多い文書ほどスコアの数値は高くなります。
- 検索文字列が文書のどの場所に含まれるかで、スコアに加算される数値が変わることがあります。登場回数が同じ 1 回でも、たとえば検索文字列がタイトルに使用されている文書は、そうでない文書よりもスコアが高くなります。

複数の検索文字列が指定された場合は、ヒットした文書数の少ない文字列をより多く含む文書の方がスコアが高くなる傾向があります。

付録 A　文書フィルタとプロパティ検索詳細

文書フィルタ

製品版の Kabayaki は、文書フィルタと呼ばれる外部プログラムと連携して、HTML やテキストファイル以外の形式の文書も検索の対象とすることができます。標準で添付されている文書フィルタは、Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF について、日本語全文検索システム Namazu の文書フィルタと比べ、より高い精度の検索を提供します。

ジャストシステム 一太郎等の文書形式にも対応します。

対応文書

- テキスト文書 JIS (ISO-2022-JP) / EUC-JP / Shift_JIS / UTF-8 / UTF-16
- HTML 文書
- Microsoft Word 95 / 97 / 98 / 2000 / 2002 (XP) / 2003 / 2007 / 2010 / 2013
- Microsoft Excel 95 / 97 / 2000 / 2002 (XP) / 2003 / 2007 / 2010 / 2013
- Microsoft PowerPoint 95 / 97 / 2000 / 2002 (XP) / 2003 / 2007 / 2010 / 2013
- Microsoft Visio 2002(XP) / 2003 / 2007 / 2010
- Microsoft Word for Mac 98 / 2001 / 2004 / 2008 / 2011 for Mac
- Microsoft Excel for Mac 98 / 2001 / 2004 / 2008 / 2011 for Mac
- Microsoft PowerPoint for Mac 98 / 2001 / 2004 / 2008 / 2011 for Mac
- Microsoft XPS 1.0
- Microsoft Rich Text Format
- JustSystems 一太郎 Ver.5-Ver13 / 2004 - 2013
- Adobe Systems Acrobat 4.0 / 5.0 / 6.0 / 7.0 / 8.0 / 9.0 / X / XI
- PDF 1.2 / 1.3 / 1.4 / 1.5 / 1.6 / 1.7
- RTF 1.0 - 1.9
- XML / SGML

その他の文書については、Namazu 2.0.21 と同等です。

XML ファイルとして Kabayaki が処理する文書は、拡張子が .xml の XML ファイルが対象で、タグと属性以外の文字列を抽出する仕様です。

処理可能な最大ファイルサイズは 100M、最大テキストサイズは 50M です。

プロパティ検索

プロパティとは、文書を識別するためにファイルに埋め込まれた文書の属性情報のことです。このプロパティを検索の対象とすることができます。プロパティには、タイトル、作成者の名前、キーワード、コメントなどを記録されています。プロパティに関しては、各文書を作成するアプリケーションの取扱説明書をご覧ください。

表 1 製品版 Kabayaki で検索できるプロパティ

プロパティ	説明	Word	Excel	PowerP	PDF	一太郎	OASYS
title	タイトル	○	○	○	○	×	×
author	著者	○	○	○	○	×	×
company	会社名	○	×	○	×	×	×

OLE オブジェクト検索

製品版の Kabayaki は、OLE で埋め込まれたオブジェクトも検索の対象とすることができます。OLE に関しては、OLE をサポートしているアプリケーションの取扱説明書をご覧ください。

表 2 製品版 Kabayaki で検索できる OLE オブジェクト

Dest/Source	Word	Excel	PowerPoint
Word	○	○	×
Excel	○	○	×
PowerPoint	×	×	×